General Insurance Seminar
Transform the Future
12-13 November 2018 • Sofitel, Sydney

Actuaries Institute

The missing link: squeezing extra performance out of predictive models

Prepared by Hugh Miller and Tom Moulder

Presented to the Actuaries Institute
General Insurance Seminar
12-13 November 2018

This paper has been prepared for the 2018 General Insurance Seminar.
The Institute's Council wishes it to be understood that opinions put forward herein are not necessarily those of the Institute and the Council is not responsible for those opinions.

© Taylor Fry

The Institute will ensure that all reproductions of the paper acknowledge the author(s) and include the above copyright statement.

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100 **f** +61 (0) 2 9239 6170
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

The missing link: squeezing extra performance out of predictive models

Hugh Miller[1] **Tom Moulder**[2]

Abstract: The 'link' function is a function that transforms a prediction from the real line to the zone of prediction. It is a core part of many modelling approaches including generalised linear models, penalised regression, tree-based ensemble prediction and neural networks. However, relatively little attention is given to the choice of link function despite evidence that they are not optimal in many cases. This is particularly important in insurance pricing contexts where there is a need to derive an accurate risk premium for customers. We discuss two more general families of link functions. We demonstrate on real and simulated data that they can help overcome situations where link functions fail, such as initial misspecification of link or the presence of overly many interactions.

1 Introduction

Model prediction is a cornerstone of modern insurance. Pricing models are used to understand the risk cost of customers based on known characteristics. Detailed claim case estimate models can help manage liabilities, particularly for long-tailed lines. And customer behaviour models can help understand how customers will interact with a company, such as accepting a quote.

Prediction is often done through regression models. Regression models and their variants now occupy a dominant position in actuarial work and predictive modelling more generally. Generalised linear models (McCullagh & Nelder, 1989) are a flexible family of models that can handle different types of responses such as Bernoulli (0-1), count and continuous responses. Penalised regression models such as the Lasso (Tibshirani, 1996, Friedman et al., 2010) extends this framework to perform simultaneous prediction and variable selection. Generalised additive models (Hastie & Tibshirani, 1990) extend the linear model to introduce nonlinearities through the controlled introduction of cubic splines.

All these models take the form:

$$y_i = g^{-1}\left(\beta_0 + x_i^T \beta\right) + \epsilon_i$$

Where the response for the ith observation, y_i, is expressed as a function plus error ϵ_i, with zero mean. The function is a transformation of a linear combination of the predictors p-vector x_i multiplied by a corresponding coefficient $(p+1)$-vector $(\beta_0, \beta_1, \dots, \beta_p)$. The inverse link function g^{-1}, among other things, takes the prediction from the real line to a domain relevant to the response. For example, a log-link means that g^{-1} is the exponential function. This ensures predictions are positive, which is important for modelling some response variables such as expected claims frequency or cost.

Other types of models, such as neural nets and boosted decision tree regression, may also make use of link functions, and many of our comments on link function choice are relevant to them too.

[1] Taylor Fry. Level 22, 45 Clarance St Sydney 2000. hugh.miller@taylorfry.com.au

[2] Taylor Fry. Level 22, 45 Clarance St Sydney 2000. tom.moulder@taylorfry.com.au

Much attention is typically given to the choice of the predictors x_i, and their related coefficient estimates. Considerations include whether:

- to find additional predictors to add to a model (e.g. add new geographic variables to a claims model)
- to expand the use of splines or polynomial terms to increase accuracy (such as generalised additive models), or perform other types of pre-processing on the input predictors
- biasing or penalising parameter estimates for variables improves accuracy (such as the Lasso)
- introducing interaction terms (one variable multiplied by another) improves the model.

This attention is justified; choosing good predictors is the most important element of a strongly performing model.

Similarly, significant attention is usually given to the error distribution, characterised by ϵ_i. The optimisation of coefficients is typically done assuming an error distribution (sometimes implicitly). Most textbooks and courses teach diagnostics for checking the distributional assumptions of a model and discuss the implications of these assumptions failing.

Relatively less attention is paid to the link function $g(\cdot)$. Virtually all models the authors have seen use 'standard' choices of link functions such as the identity, log, inverse, logit or probit functions. We believe that this practice is a combination of:

- Tradition – these links have a long history of use
- Convenience – most statistical software has the standard link functions built in
- Intuition – The log-link in particular is popular as it converts an additive model into a multiplicative one ('males at 10% to the predicted cost'). The logit link has a similar intuition around the log-odds ratio.

This list does not include accuracy. For most real-world datasets there is no reason to suppose that a standard link is 'optimal'. The link function itself also imposes strong assumptions on how a variable relates to the response over the entire range of predicted values.

Through real-world and simulated datasets, we argue that predictions can be improved by paying more attention to link functions.

There is limited existing literature that discusses link functions that we are aware of. McCullagh & Nelder (1989) discuss the notion of the canonical link, which is one further argument for picking a 'standard' choice. Mallick & Gelfand (1994) discuss links based on mixtures-of-beta distribution cumulative density functions, and Bayesian approaches to fitting. Some other relevant work is by Czado (1992), Czado & Raftery (2006) and Flach (2014); one of the families we explore is based on the latter. Miller (2017) explores alternative link function choices in an insurance reserving context.

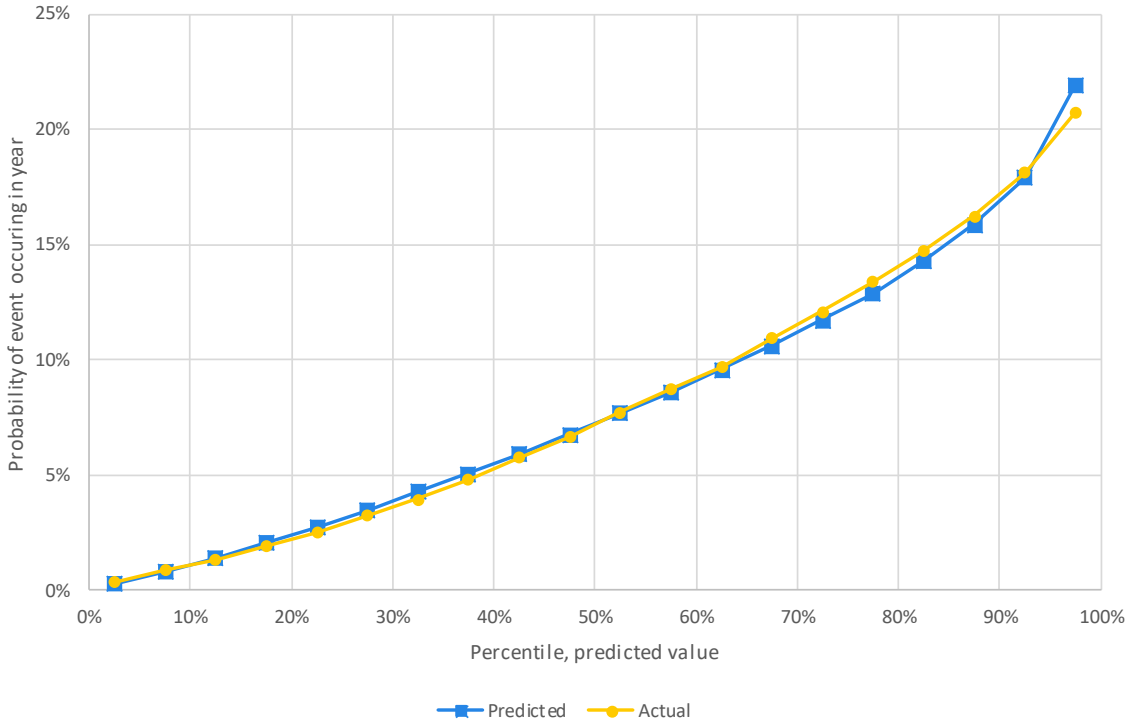
2 Motivation for change – visible model misfits

A common diagnostic for a fit is the actual versus expected by predicted percentile (AvEPP) plot. This:

- Sorts the dataset from lowest to highest predicted value
- Divides into equally sized groups (20 groups, say, with 5% of observations in each)
- Calculate and plots the average predicted and the average actual response for each group.

This plot is useful since it gives both a sense of the spread of predictions, and how well this spread reflects the actual spread (as detectable by the model). Figure 1 shows such a plot for a probability model (0-1 response) with logit link. While the range of predicted values is large, there are some misfits visible. The top-most 5% of predictions overestimate the response, with an offsetting underestimate in the 70-90 percentile range.

Figure 1 Example AvEPP plot for real 0-1 regression model, logit link



It is possible to add confidence intervals to the plot, and other rigour (such as 'run' tests to check for unusually high or low set of predictors).

A misfit in an insurance pricing model can have significant implications. It implies a level of cross-subsidisation in the fit that potentially overstates risk cost for some customers and understates it for others. Insurers would want a better fitting model to remove hidden cross-subsidies when estimating risk cost.

A misfit on the diagnostic is traditionally interpreted as either:

- A poor set of terms related to a continuous variable. For example, extrapolating a trend at the extreme of an age curve may produce over or under-fit.
- An absence of necessary interactions.

These interpretations are true; the addition of enough continuous splines and interactions should always be enough to correct an AvEPP plot misfit. This can be hard in practice. The size of the dataset may limit the number of terms (splines or interactions) that can be added. Or modellers are faced with a trade-off between a term that improves the chart but is not statistically significant on other tests. Correcting misfits can also be a time-consuming trial and error process.

However, there is a third potential interpretation of an AvEPP plot misfit – the need for a different link function. An altered link function will change the predicted values directly and potentially negate the need for additional terms and interactions. This can be seen easily – take a model with a good fit and change the link (from a log to an identity, say), and it will usually produce effects of the type seen in the figure above.

We are particularly interested in misfit at the extremes. We regularly observe patterns on real datasets that either undershoot or overshoot at the extreme left or right of a plot, which then induces corresponding biases elsewhere on the AvEPP curve. This is suggestive of effects that could be at play on a dataset:

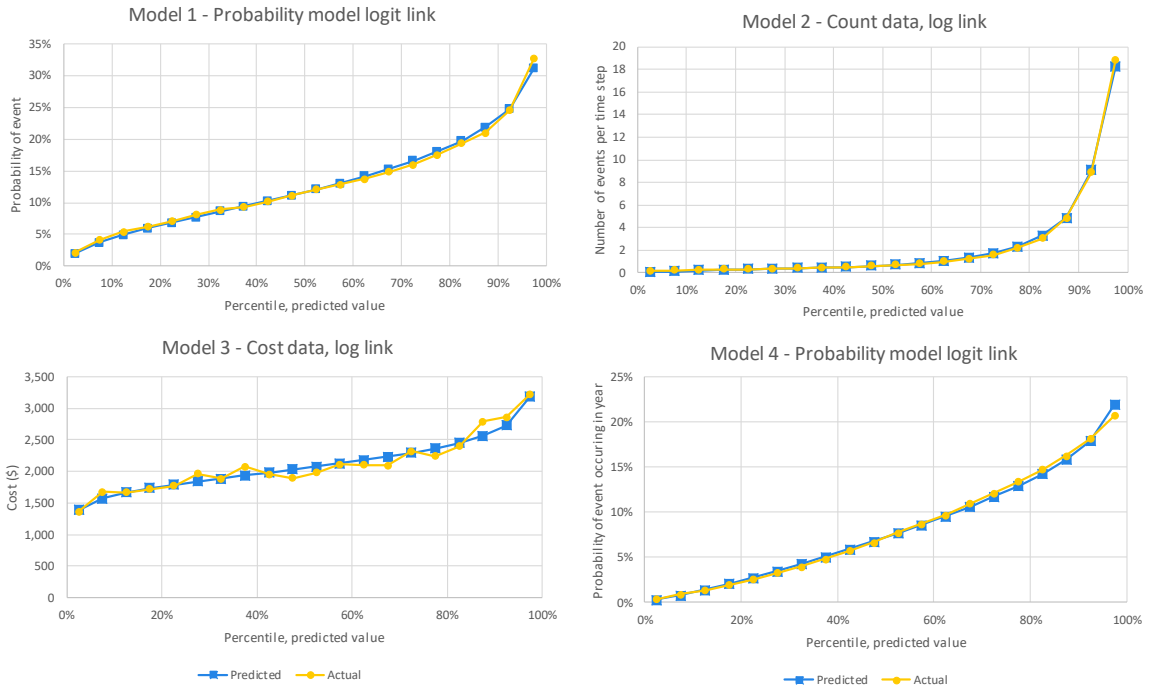
» **Diminishing returns at extreme.** It may be that for predictions that are already high, the impact of a factor is less than at more moderate values. Young age might increase average claim cost, but the impact of age might be less if expected claim cost is already high. Figure 1 is of this type.

» **Extreme concentration**. Opposite to the previous bullet, sometimes the highest risk groups are 'off the scale' in terms of response relative to effects defined of the broader population. While the model might identify the subgroup, they may undershoot the actual effect.

Note that these observations also suggest that optimal link functions cannot be pre-determined; if adjustments to the curve are required, it will very much depend on the level of the predictions for that model.

The figure below shows four AvEPP plots for different types of models. Model 4 is the same as Figure 1. All show some form of misfit.

Figure 2 Example AvEPP plots based on four real-world datasets, various setups



Bias can be seen in the AvEPP plots for each of the models shown above.

» In model 1, predictions are visibly below actual values for the highest prediction band. Offsetting this, predictions are biased above actual values between the 60th and 95th percentile of predictions. This is consistent with an 'extreme concentration' effect.

» In model 2, a count data (Poisson) prediction, the visible plot looks reasonable, with perhaps slight underprediction at the top (rightmost) end. However, there is significant relative bias at the lower predictions, with predictions being about half of actual over the 0-30th percentiles. This is consistent with a situation that has a base level risk that sits slightly above zero. The bias is visible at the low predicted values when viewed on a log scale, which we show in Figure 12 in Section 4.7.

» In model 3, predictions appear below actual values above the 85th percentile of predictions. We observe offsetting higher than actual predictions between the 40th and 85th percentile of predictions.

» In model 4, which is a reproduction of Figure 1, predictions are greater than actuals both for the highest predictions, and between the 20th and 40th percentiles of predicted values. Offsetting this, predictions are below actual values between the 65th and 90th percentile of predictions. This is consistent 'diminishing returns at extreme' effect.

When an analyst observes such misfits, they must be either accepted or remedied. Post-model processing can address the misfit (for example, a separate calibration stage). This is direct and relatively flexible; however, this creates an additional layer of complexity to the model fitting process and must be regularly updated. It is also often performed by grouping predictions into buckets prior to calibration, which can result in non-smooth effects in the final model. Improving the model directly would be preferable.

3 Families of link functions and implementation

We have tested two approaches creating flexible link functions, that can correct for the model misfits described in Section 2.

3.1 Generalised link functions

We propose two families of link functions to extend the logit and log transformations. They are 3-parameter families of transformations that appear to have good performance on a range of examples.

Our generalised logit link function, applied to the linear predictor LP, is:

$$g^{-1}(LP) = B + \frac{1}{(A + \exp(-(LP - C)))^K}$$

For $A = 1$, $B = 0$, $C = 0$ and $K = 1$ we recover the usual logit link. We do not count C as a parameter since it simply shifts the linear predictor.

Similarly, our generalised log link function is

$$g^{-1}(LP) = \frac{\{A + \exp(LP - C)\}^K - A}{K} + B$$

For $A = 1$, $B = 0$, $C = 0$ and $K = 1$ we recover the usual log link. As K approaches zero the function approaches a softmax function. We do not count C as a parameter since it simply shifts the linear predictor; however it is needed for practical optimisation as it is needed to balance out changes in K.

The parameters A, B, and K could in principle be optimised simultaneously in a maximum likelihood exercise. We have used an iterative procedure (see Section 3.3) where we switch between a normal regression (with current link function) followed by a link parameter optimisation step. We repeat until convergence in the maximum likelihood.

Since the impacts of the new parameters are not linear in the linear predictor space, the procedure introduces the risk of finding local optimisers rather than global solutions. Our approach is to start with parameters set at the 'base' link (the logit or log). If we do recover a local solution, it will at least have the virtue of being 'close' to the standard regression fit.

3.2 Pre-link linear predictor modifications

We have also experimented with the parametric link functions specified in Czado (2007) and developed in Flach (2014). This approach applies makes modifications to the model output on the linear predictor scale, before the standard link function is applied.

The modifications take the form of power transformation functions.

Right tail modification

$$h(LP,\psi_1)=\begin{cases} LP_0+\ln(LP-LP_0+1), & LP\geq LP_0 \text{ and } \psi_1=0 \\ LP_0+\dfrac{(LP-LP_0+1)^{\psi_1}-1}{\psi_1}, & LP\geq LP_0 \text{ and } \psi_1\neq 0 \\ LP, & Otherwise \end{cases}$$

Left tail modification

$$h(LP,\psi_2)=\begin{cases} LP_0-\ln(-LP+LP_0+1), & LP\leq LP_0 \text{ and } \psi_2=0 \\ LP_0-\dfrac{(-LP+LP_0+1)^{\psi_2}-1}{\psi_2}, & LP\leq LP_0 \text{ and } \psi_2\neq 0 \\ LP, & Otherwise \end{cases}$$

When $\psi_1=1$ and ψ_2=1 then the linear predictor is unchanged. Values larger than one accelerate the linear predictor (it grows faster than usual), and lower values will taper it; values less than 0 introduce asymptotes in the modified predictor.

We allow both left and right tail modifications, and extend the approach of Flach (2014) in the following ways:

- Allowing both left and right modifications with a (variable) gap between them. LP_0 is therefore replaced by LP_1 in the right tail modification and LP_2 in the left, with $LP_2\leq LP_1$.
- Optimising the parameters of the link function modification simultaneously with the selection of GLM coefficients. Other attempts (Flach, 2014) pre-define the link modifications in advance, rather than tailoring to the dataset.

We can specify modifications as a four-parameter modification to the link function.

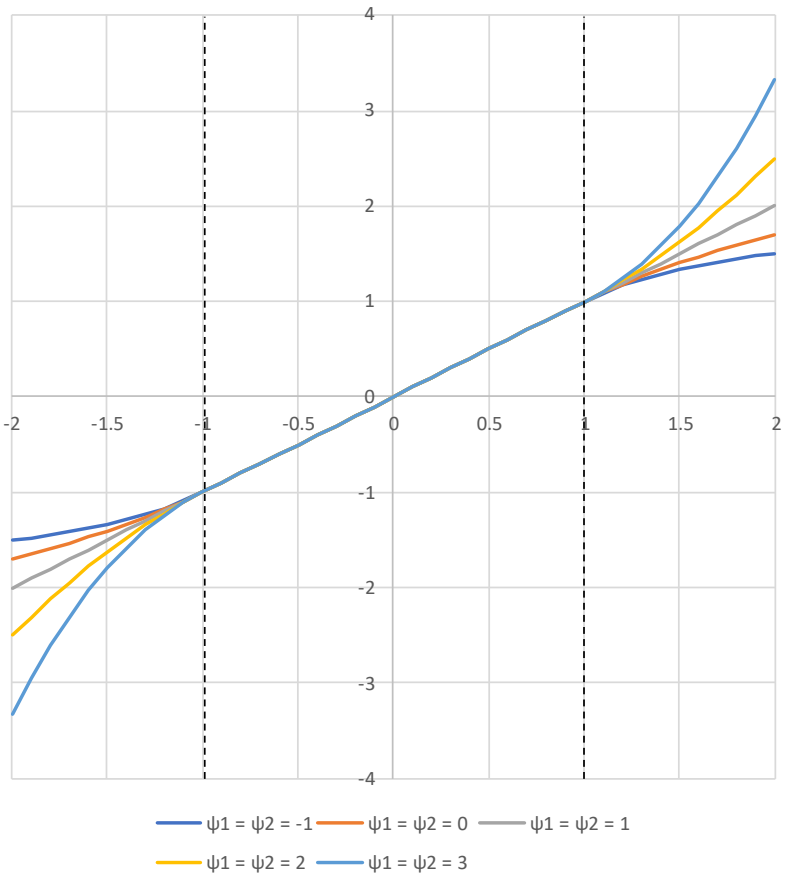
Two-sided tail modification

$$h(LP,\psi_1,\psi_2,LP_1,LP_2)=\begin{cases} LP_1+\ln(LP-LP_1+1), & LP\geq LP_1 \text{ and } \psi_1=0 \\ LP_1+\dfrac{(LP-LP_1+1)^{\psi_1}-1}{\psi_1}, & LP\geq LP_1 \text{ and } \psi_1\neq 0 \\ LP_2-\ln(-LP+LP_2+1), & LP\leq LP_2 \text{ and } \psi_2=0 \\ LP_2-\dfrac{(-LP+LP_2+1)^{\psi_2}-1}{\psi_2}, & LP\leq LP_2 \text{ and } \psi_2\neq 0 \\ LP, & Otherwise \end{cases}$$

Because the modification is made before the application of a standard link function, the method applies more generally than the modifications in section 3.1, including all common links like the identity, log and logit link functions.

The impact of tail modifications on the linear predictor is shown in the figure below.

Figure 3 Impact of tail modification, $LP_2 = -1$, LP_1=1.



3.3 Optimisation and implementation

Irrespective of which of the approaches outlined in 3.1 or 3.2 is followed, our approach is to integrate link function selection as part of the model fitting process. We:

» Maximise the log-likelihood under the assumed error distribution (as per standard regression modelling). Or equivalently, minimise the deviance.

» Iterate between two optimisations steps until convergence:

1. Estimation of linear predictor coefficients (a standard regression fit) to maximise log-likelihood, holding the link function constant
2. Estimation of link function parameters, holding the linear predictor coefficients constant

We initiate the fit using a 'standard' link function.

This approach was implemented in R using a combination the *glm* package and non-linear optimisation function *optim*. The *optim* package was used for choosing parameters for the link function, and in some cases was also used to select predictor coefficients (rather than the standard fitting solutions in the *glm* package).

4 Theoretical setup, simulated examples and other results

4.1 Simple simulation example

As a simple test of the approach, we developed a two-variable example problem using simulated data. We simulate 10,000 observations of data, using the following parameters:

$$\mu_i = 0.05x_I + 2.5x_2 - 0.025x_1x_2$$
$$y_i \sim Poisson(u_i)$$
$$x_1 \in [0, 100], x_2 \in \{0, 1\}$$

We then fit a standard Poisson/log link GLM with no interaction terms and tested the improvement in fit obtained by the modifications introduced in Section 3.

The fit from the standard GLM is poor for two reasons:

- The GLM does not contain an interaction term, despite the data having a strong negative interaction between x_I and x_2
- The simulated effect of the predictors on the mean is linear (that is, the 'true' link function is linear). However, the application of the log link function means that the default modelled effect is exponential.

The revised fit from apply the generalised link function approach is shown in Figure 4.

Applying the generalised log link function improves the fit for observations in the bottom 85% of predictions. Interestingly, the approach failed to fully correct for the misfit at the top end of observations. This could be due to either insufficient flexibility of the link function to address such a severe misfit, or failure of the optimisation to converge to a globally optimal parameterisation.

Figure 4 Simple example, base GLM and improved fit using generalised log fit[3]

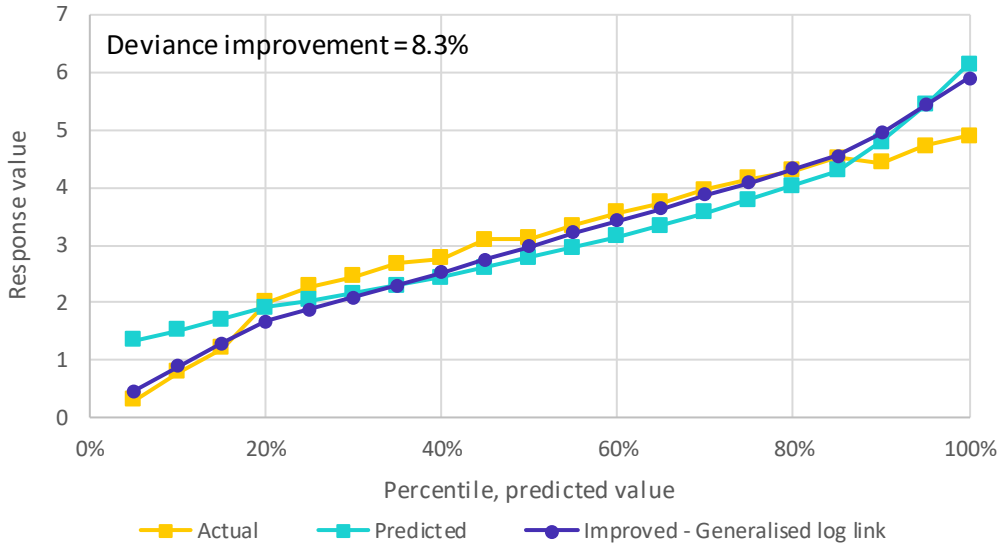
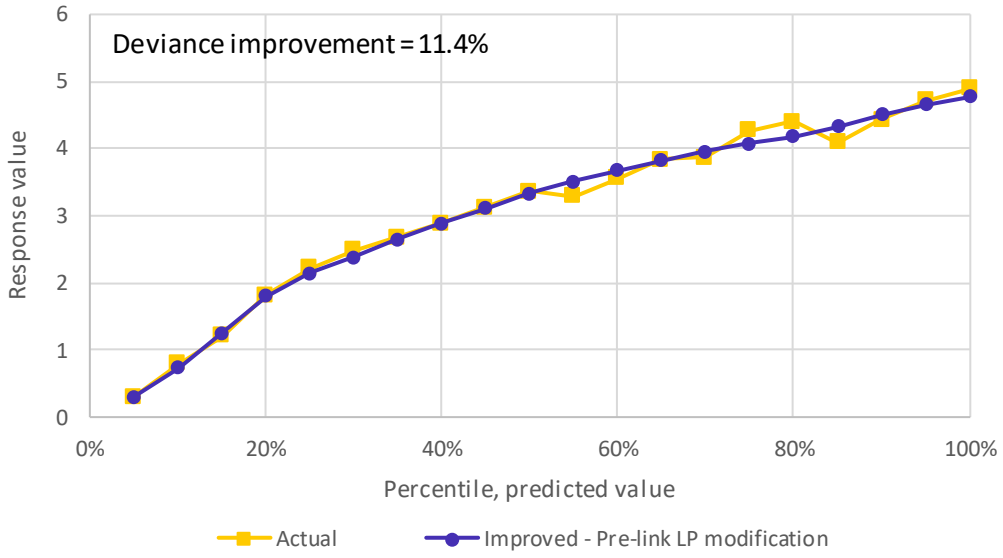


Figure 5 shows the AvEPP for the fit using a pre-link linear predictor modification. It is separated because the linear predictor modification process allows GLM coefficients to update, which means that both the actual and predicted curves change.

Applying pre-linear linear predictor modifications significantly improved the fit over the full range of predictions.

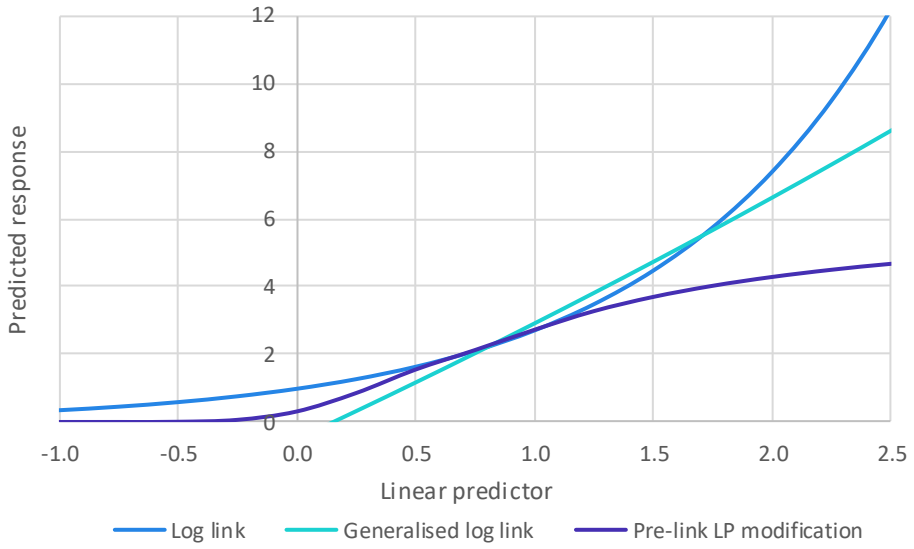
Figure 5 Simple example, pre-link linear predictor modification fit



[3] The generalised log and logit fit charts are presented using only a single link function optimisation (that is, no re-estimation of β coefficients). This was partly for practical reasons; first, in cases where negative values were predicted the further optimisation becomes difficult. Second, the single step allows us to present old and new fits on the same AvEPP plot, since there is no reordering of prediction.

Figure 6 shows the optimised link function shapes that underlie the fits shown in Figure 4 and Figure 5.

Figure 6 Simple example, selected inverse link functions after optimisation



» The generalised log link approach has adopted an almost linear link function. This is intuitively reasonable, as negative interaction in the data incentivises the optimisation to dampen the curve of the logistic function at both the top and bottom end of predictions. Note that in this case it also allows negative predictions for the modified link.

» The pre-link LP modification approach has also adopted a link function with downward curvature at both the left and right tails. Again, this is consistent with what we'd expect for data containing a negative interaction.

4.2 More general simulation setup

We have interpreted the concepts of 'Diminishing returns at extreme' and 'Extreme concentration' from Section 2 in a way amenable for simulation.

Suppose we have a linear predictor defined by main effects and pairwise interactions:

$$\eta_i = \beta_0 + \sum_j \beta_j x_{ij} + \sum_{j,k} \gamma_{jk} x_{ij} x_{ik}$$

Where:

» The x_{ij} are all Normally distributed with mean zero and variance one – $N(1,1)$.

» The β_j are sampled from a modified double exponential distribution where $|\beta_j| \sim \exp(\phi_1)$ and $P(\beta_j > 0) = \rho_1$. Typically, we take $\rho_1 > 0.5$ so the parameters are more likely to be positive.

» The γ_{jk} are sampled from a modified zero-inflated double exponential distribution where $P(\gamma_{jk} = 0) = \delta$, but conditional on being nonzero $|\gamma_{jk}| \sim \exp(\phi_2)$ and $P(\gamma_{jk} > 0) = \rho_2$.

The setup means that there can be many interactions, many of which will be difficult to detect in a standard regression. If we take $\rho_2 < 0.5$, then more correlations are negative; and this will tend to show diminishing returns at extreme. Similarly, $\rho_2 > 0.5$ will correspond to extreme concentration type behaviour.

We've generally looked at the improvement in training set deviance and used datasets sufficiently large so that overfitting effects are less relevant. Nonetheless, we've also tracked deviance improvements on out of sample data, to test the performance of the modifications when generalising to unseen data.

4.3 Simulation experiment 1

Experiment 1 tests the generalised log link function on simulated data, in the case where main effects β_j are positive, but interactions γ_{jk} are mostly negative. This corresponds to diminishing returns at the extreme, as described Section 2.

We simulated data on the linear predictor scale using the process described in Section 4.1, and the following parameters: $\rho_1 = 1,\ \rho_2 = 0.2$, ϕ_1 = 0.1. ϕ_2 = 0.1, δ = 0.5. Setting $\rho_2 < 0.5$ ensures that interactions are negative on average. A simulated target variable is then created for each observation i by sampling from $Poisson(e^{\eta_i})$. We used 10,000 sample observations, and 10 predictor variables.

We then constructed a Poisson main effects GLM model on the simulated data, and applied both the generalised log-link function described in Section 3.1, and the pre-link LP modification described in Section 3.2.

Results of this experiment are shown in Figure 6 (one simulation) and Figure 7 (100 simulations) below. The AvEPP plots for the results from the pre-link LP modification approach are shown separately, because re-estimating regression coefficients allows the shape of the Actual curve to change.

Figure 6 Experiment 1, base GLM and improved fits using generalised log, and pre-link LP modification

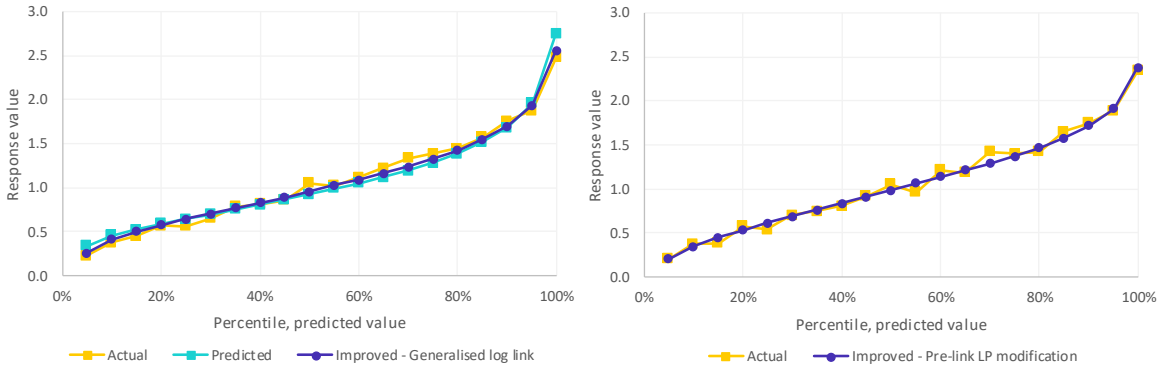


Figure 6 shows an example where the base GLM over-predicts in the bottom 10% and top 5% of the data. This is offset by under-predictions between the 40th and 80th percentiles of the data, giving the characteristic shape associated with missing negative interactions.

- Applying the generalised log link function visually improves the fit across the range of the response.
- The pre-link LP modification approach produces an AvEPP plot that fits well across the full range of responses.

Figure 7 Experiment 1, base GLM and improved fits. Average over 100 simulations

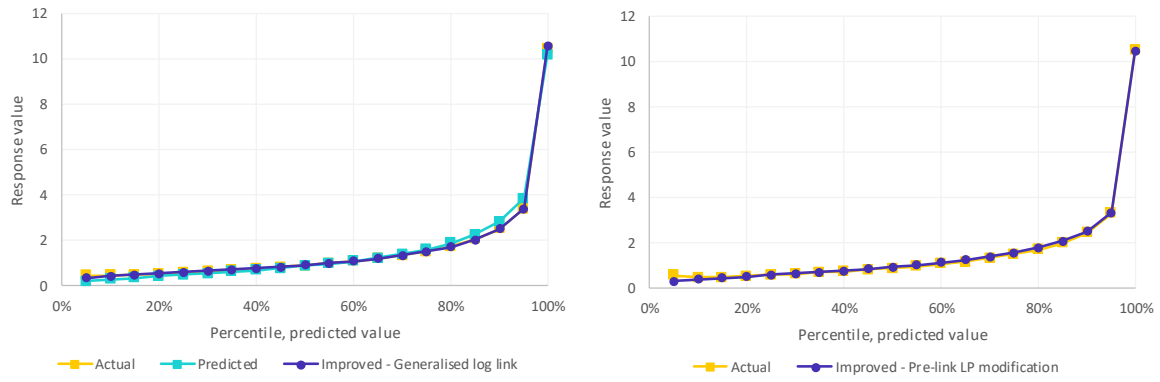


Table 1 Average deviance improvement over 100 simulations on out of sample test data

Approach	Average test set deviance improvement
Generalised link	5.7%
Pre-link LP modifications	6.0%

- Figure 7 show that on average, base GLM predictions (light blue, left pnael) are too low for the bottom 30% of observations, and the highest 5%. Offsetting these underpredictions, the base GLM over-predicts between the 60th and 95th percentiles of the data.
- Both the generalised link function and pre-link LP modification approaches visually improve the fit on average, and remove the bias visible in the base GLM.
- Deviance improved on out of sample data for under both approaches. On average, the improvement was slightly better (6.0%) under the pre-link LP modification approach than the generalised link approach (5.7%).

4.4 Simulation experiment 2

Experiment 2 tests the generalised logit link function on simulated data, in the case where main effects β_j are positive, but interactions γ_{jk} are generally negative.

Data was simulated using the same parameters described for Experiment 1 in Section 4.3, except the simulated response is sampled from a Bernoulli distribution with mean $logit^{-1}(\eta_i)$. We also set $\beta_0 = -2$ to scale the repose to an appropriate range (an average response rate of about 20%), and use a sample size of 100,000.

We then constructed a Binomial main effects GLM model on the simulated data and applied the link function modifications approaches described in Section 3.1 and Section 3.2. Results of this experiment are shown in and Figure 8 and Table 2 below.

Figure 8 Experiment 2, base GLM and improved fits. Average over 100 simulations

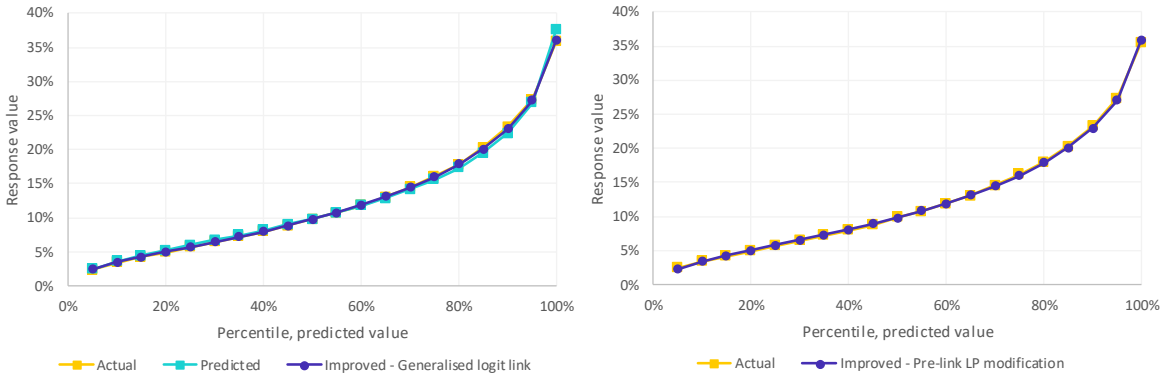


Table 2 Average deviance improvement over 100 simulations on out of sample test data

Approach	Average test set deviance improvement
Generalised link	0.24%
Pre-link LP modifications	0.28%

- On average, base GLM predictions are slightly too high for the bottom 45% of observations, and the highest 5% (light blue line, left panel). Offsetting these overpredictions, the base GLM under-predicts between the 60th and 95th percentiles of the data.
- Both the generalised link function and pre-link LP modification approaches visually improve the fit on average, and remove the bias visible in the base GLM.
- Deviance improved on out of sample data for under both approaches. On average, the improvement was slightly better (0.28%) under the pre-link LP modification approach than the generalised link approach (0.24%).
- Due to the difference in data and distribution assumptions, deviance improvements should not be compared with those from other experiments.

4.5 Simulation experiment 3

For experiment 3, we return to the Poisson distribution, log-link, and replace $\rho_2 = 0.8$, so that interactions are more likely to be positive and compound main effects. We also set $\phi_1 = 0.05$ and $\phi_2 = 0.05$ to keep predictions in a reasonable range. All other parameters are unchanged from experiment 1.

The results of this experiment averaged over 100 simulations are shown in Figure 9 and Table 3.

Figure 9 Experiment 3, base GLM and improved fit using generalised log fit

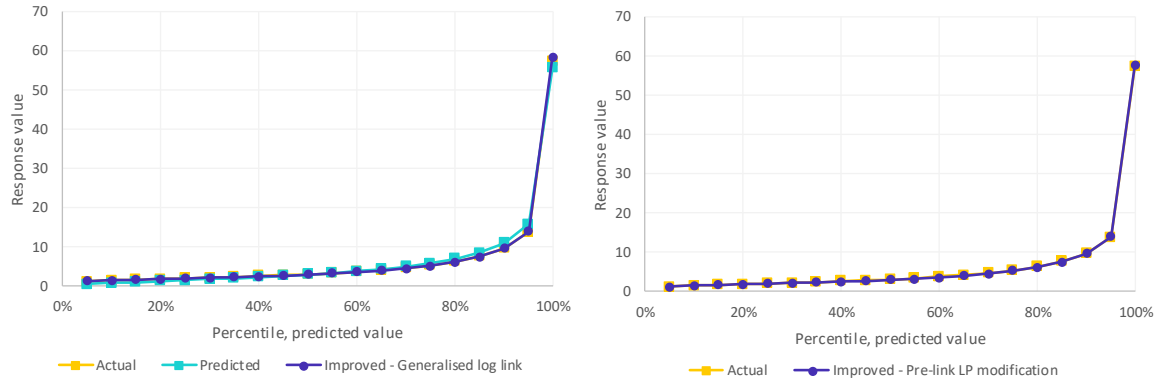


Table 3 Average deviance improvement over 100 simulations on out of sample test data

Approach	Average test set deviance improvement
Generalised link	12.5%
Pre-link LP modifications	12.7%

- Again, bias is visible in the base GLM predictions across the AvEPP curve, with underpredictions at the top and bottom end of the response, offset by overpredictions through the middle.
- Both the generalised link function and pre-link LP modification approaches visually improve the fit on average, and remove the bias visible in the base GLM.
- Deviance improved on out of sample data for under both approaches. On average, the improvement was slightly better (12.7%) under the pre-link LP modification approach than the generalised link approach (12.5%).

4.6 Simulation experiment 4

For experiment 4, we return to a Binomial model with a logit link, setting $\rho_2 = 0.8$, so that interactions are more likely to be positive and compound main effects. We also set $\phi_1 = 0.05$, $\phi_2 = 0.025$ and $\beta_0 = -3$ to keep predictions in a reasonable range. All other parameters are unchanged from experiment 2.

The results of this experiment averaged over 100 simulations are shown in Figure 10 and Table 4.

Figure 10 Experiment 3, base GLM and improved fit using generalised log fit

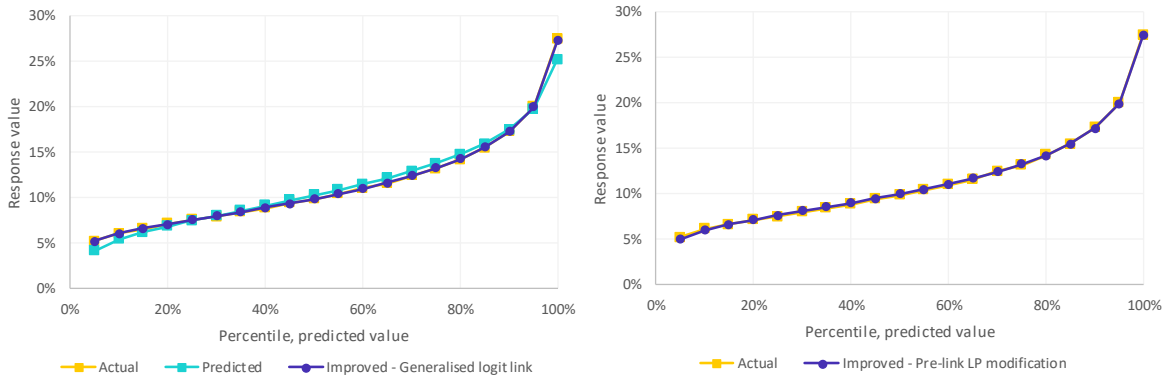


Table 4 Average deviance improvement over 100 simulations on out of sample test data

Approach	Average test set deviance improvement
Generalised link	0.08%
Pre-link LP modifications	0.07%

- Again, bias is in the base GLM predictions across the AvEPP curve, with underpredictions for the bottom 15% and top 5% of the response. This is offset by overpredictions between the 35th and 85th percentiles.
- Both the generalised link function and pre-link LP modification approaches visually improve the fit on average, and remove the bias visible in the base GLM.
- Deviance improved on out of sample data for under both approaches. On average, the improvement was slightly better (0.08%) under the generalised link approach than the pre-link LP modification (0.07%).

4.7 Real-world examples

We've used the generalised logit and log curves (section 3.1) to do a one-step optimisation of the real-world examples introduced in Section 2. Figures for models 1, 2 and 4 are shown below, followed by commentary.

Figure 11 Model 1, improved fit using generalised logit fit. (A, B, C, K) = (1.0054, -0.00194, -0.797, 1.935)

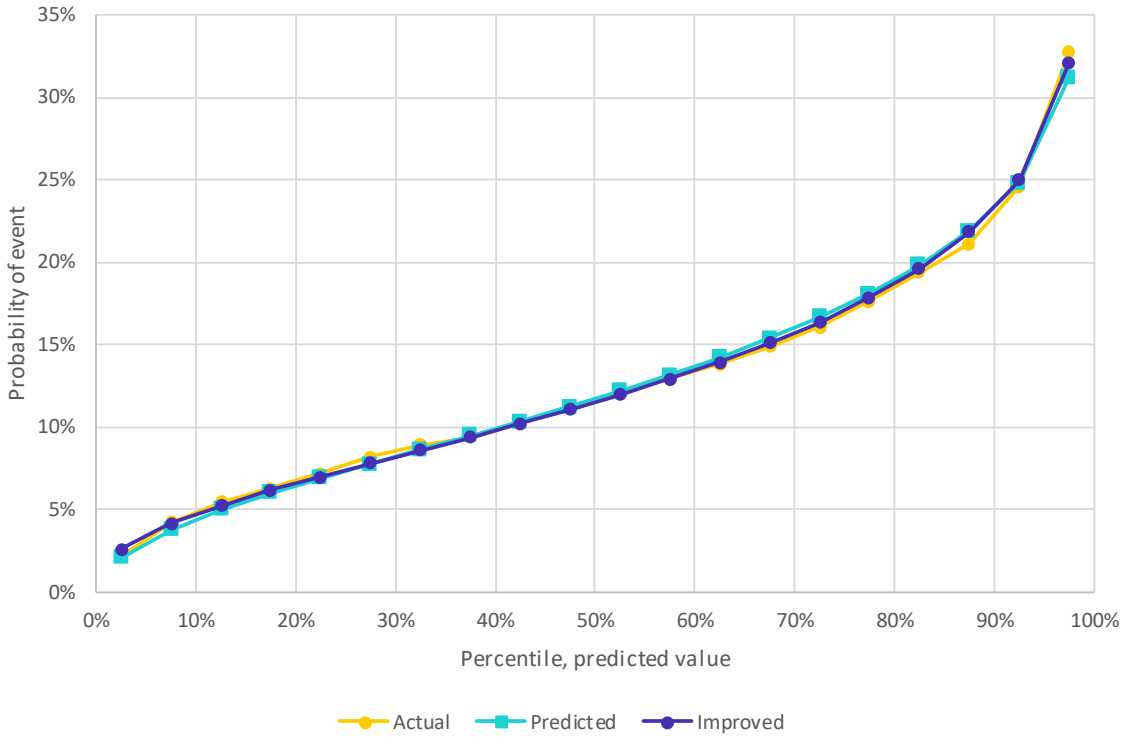


Figure 12 Model 2, improved fit using generalised log link. (A,B,C,K) = (-0.00796, 0.0627, 0.0642, 1.056)

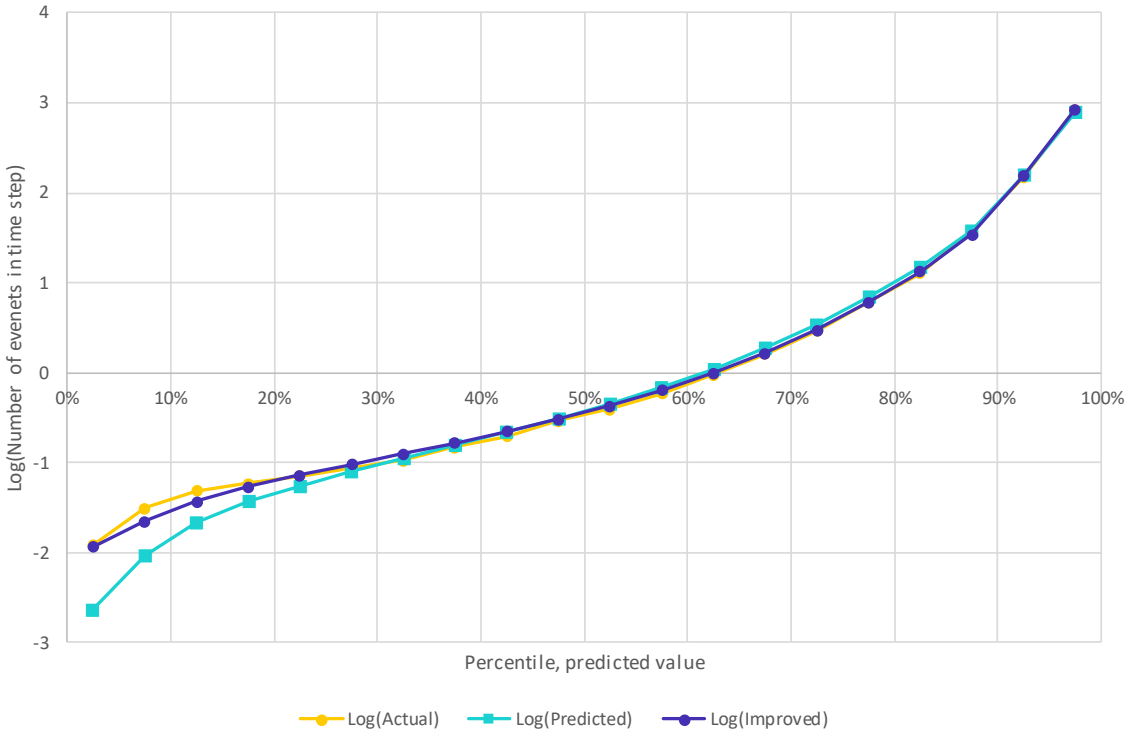
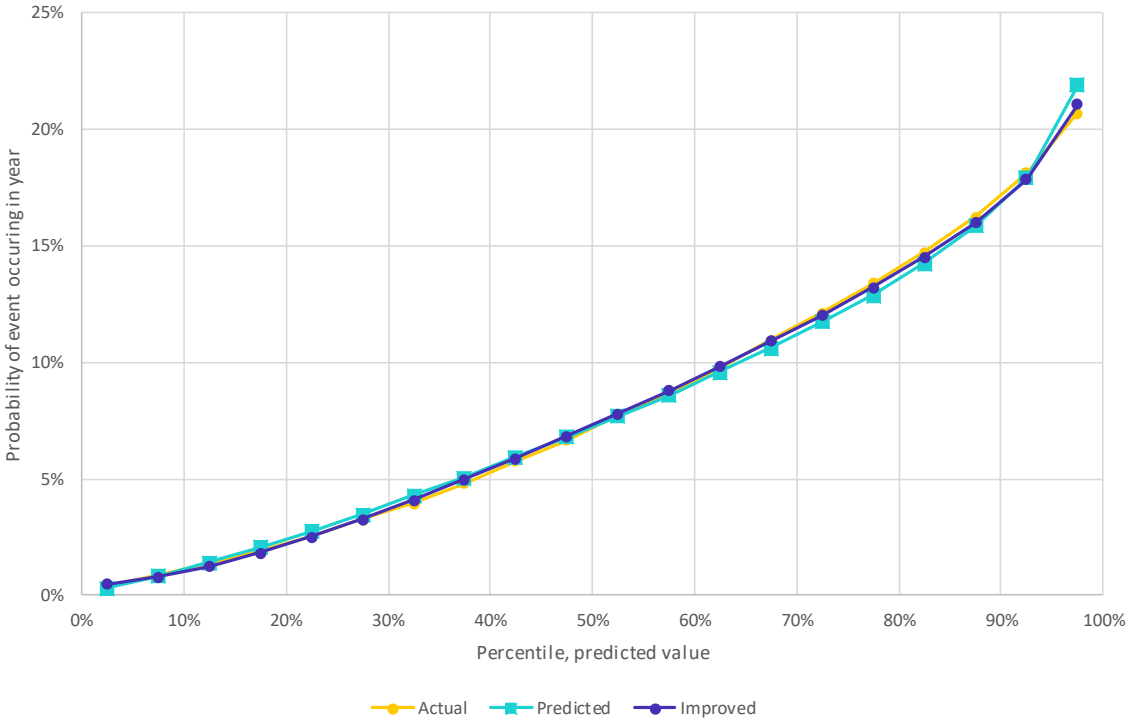


Figure 13 Model 4, improved fit using generalised logit fit. (A, B, C, K) = (1.20, 0.00847, -0.680, 1.39)



» The fit for Model 1 visibly improves the top end (highest predictions) and reduces the bias in the mid-range.

- We've shown the fit for Model 2 on the log-scale to highlight the improved fit at the low-end. The bias, while small in absolute terms, has been significantly reduced in relative terms.
- Model 3 (not shown) has no visible impact on the plotted fit – monotonic changes are insufficient to materially improve the shape of the fit, but is good validation that a 'null' result is possible, with minimal impact on predictions. The parameters for the curve were (A, B, C, K) = (-22.1, 82.1, -3.31, 0.684). Interestingly, while the optimised curve produces very similar fits on the observed range, the fit is actually flatter beyond the observed range of predictions, so outliers are likely to be less extreme in prediction.
- Model 4, similar to model 1 shows material improvements at the top end, with bias reduction in the mid-range.

The results suggest clear improvements are possible, particularly when the misfit is most evident at the upper or lower ends of the predictions.

5 Conclusions

We have demonstrated that regression model misfits can be significantly improved by adding flexibility to the link functions. Such improvements are needed when there is initial doubt around the true link function, or whether the interplay of predictor variables means than an unreasonably high number of interactions would be required to address the issue.

We have also introduced two link function families for performing such improvements. The generalised curves in Section 3.1 allow transformations of log and logit links that seem to bear resemblance to the types of issues seen on real datasets. The pre-link modifications of Section 3.2 are more geared towards finding misfits in the tails, and in some cases appear more flexible. Both approaches have been found to be effective in our explorations, although reported results focus on the generalised link functions.

There are avenues for further research in this area. Our experiments have focused on traditional generalised linear models, where the potential gains are perhaps larger than more complex machine learning models. However, there is still likely to be some value in these types of approaches in other model types, such as gradient boosted machines or neural net models; such models can exhibit similar types of issues on model diagnostic plots.

6 References

Czado, C. (1992). On link selection in generalized linear models. In *Advances in GLIM and Statistical Modelling* (pp. 60-65). Springer, New York, NY.

Czado, C., & Raftery, A. E. (2006). Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors. *Statistical Papers*, *47*(3), 419-442.

Czado, C. (2007). Fitting Generalized Linear Models with Parametric Link in S. *Technical report, York University*, North York, Ontario, Canada.

Flach, N. (2014). Generalized linear models with parametric link families in R. Bachelor's Thesis, Technische Universität München

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*, volume 43 of Monographs on Statistics and Applied Probability.

Mallick, B. K., & Gelfand, A. E. (1994). Generalized linear models with unknown link functions. *Biometrika*, *81*(2), 237-245.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC press.

Miller, H. (2017). Building better PPAC Models. *Institute of Actuaries Australia, Injury & Disability Schemes Seminar 2017, Brisbane.*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.